July 1975

B75-10140

NASA TECH BRIEF

Ames Research Center

Method of Identifying Clusters Representing Statistical Dependencies in Multivariate Data

The analysis and interpretation of environmental and medical data often is difficult because there are many simultaneously interacting variables and because there seldom is an explicit theory that suggests a relationship between observables and quantities of interest. Classical multivariate statistical techniques, such as multiple regression and principal component analysis, may provide a satisfactory analysis and interpretation when data can satisfy the rather stringent linearity requirements imposed by the techniques; however, the classical linear methods are inappropriate when the underlying relationships between variables are highly nonlinear.

Cluster analysis is a multivariate technique for detection of nonlinear relationships in data; it seeks to find, from the structure of the data itself, natural groupings or categories in the multivariate observations made on a set of objects. Unfortunately, most algorithms for cluster analysis provide no information that can be used to distinguish predictive clusters from the clusters resulting from statistical fluctuations in the data.

There has been developed a new method that combines cluster analysis with statistical tests of significance by Monte Carlo sampling procedures to detect the presence of statistical relationships among variables in multidimensional data. The basic hypothesis is that the population marginal distributions of the variables are statistically independent; assuming that there are sufficient data to be assured that the sample marginals are representative of population marginals, independent computer sampling of the marginals is used to determine how often, by chance, the generated samples show cluster structure similar to that of the data.

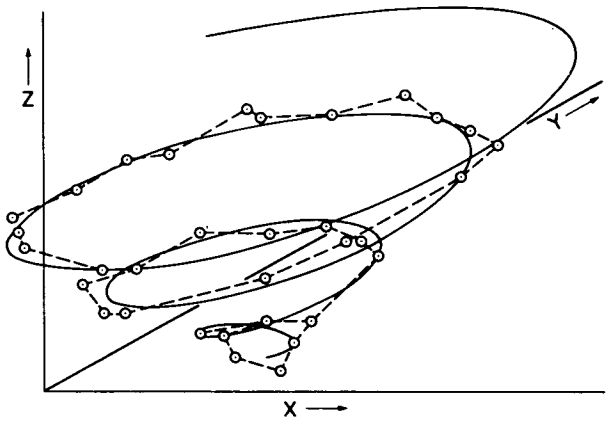
The new method is comprised of two parts: (1) an "extended group clustering algorithm" that finds compact clusters and then merges them to form extended clusters; and (2) a "cluster classification algorithm" that determines whether the clusters found by the extended group clustering algorithm could have arisen by chance from independent marginal distributions of the variables. After clustering has been accomplished, the final testing of clusters utilizes a Monte Carlo method of scrambling the data by sampling independently from the marginal distribution of variables.

The approach is first to cluster and then to compute spatial boundaries for the resulting clusters; the next step is to compute, from the set of Monte Carlo samples obtained from scrambled data, estimates of the probabilities of obtaining at least as many points within the boundaries as were actually observed in the original data. Clusters with low probability are retained for further examination because these clusters may represent statistical association between variables.

In biological systems, it is quite common to find many simultaneously interacting variables. Analysis of data from biological systems is often begun by plotting one variable against another. By using video outputs on modern computer systems, the technique can be extended to pseudo-three- or pseudo-four-dimensional displays of groups of variables; nevertheless, it cannot be expected that three- or four-dimensional displays will always meet the requirements for finding multidimensional statistical relationships in complex multivariate data. To illustrate this difficulty and the use of the present technique to discover relationships in three dimensions that are not obvious in two-dimensional plots, an artificial data set was generated to form three scattergrams,

(continued overleaf)

each of two dimensions. Visual examination of the plots did not reveal any strong relationship in the data, yet in a space of the correct number of dimensions, the relationhip becomes obvious and the data represent a noisy spiral. Random (normally distributed) noise was added to the spiral-generating function, and the final plot resulting from use of this technique is shown in the diagram.



The solid line in the diagram is the generating function, and the open symbols represent the modes found by the clustering technique. The dashed line connects modes to their two most similar neighbors, providing that the clusters of which they are members have at least one point in common. The break (just to the left of the Z-axis) in the derived curve is the result of a statistical fluctuation in the density of points, but the clustering algorithm resumed following the generating function after this break. The modes for the last 180° of the spiral are not shown because it would make the relationship between the generating function and the modes very difficult to visualize in the two-dimensional representation.

The Monte Carlo estimation of the probability of forming the first extended group at random from independent marginal distributions was less than 10^{-27}, far too small for this group reasonably to be a type II cluster. Hence, either this cluster is recognized as one representing a statistical dependence between variables (type I cluster), or it is an artifact (type III cluster). The latter possibility could be eliminated by gathering new data from the same process, and analyzing it in the same manner. If the same spiral appears, evidence becomes overwhelming that it represents real structure in the data. Even after the spiral has been detected and has been proven to be real, the task of visualizing or interpreting the relationship remains; this difficulty is always present. The new technique only searches for statistical association between variables, and can be thought of as a higher dimensional analog of the process of visual inspection of two-dimensional and three-dimensional scattergrams.

Note:

Requests for further information may be directed to:

Technology Utilization Officer
Ames Research Center
Moffett Field, California 94035
Reference: TSP 75-10140

Source: William J. Borucki,
Don H. Card, and Gilbert C. Lyle
Ames Research Center
(ARC-10744)